JOURNAL OF MULTIVARIATE ANALYSIS 37, 85–103 (1991)

Optimal Parameter Choice for Error Minimization in Bivariate Histograms

J. A. HUESEMANN AND G. R. TERRELL*

*ETH, Zürich, Switzerland; and *Department of Statistics, Virginia Polytechnic Institute, Blacksburg, Virginia 24061*

Communicated by the Editors

Scott, Freedman, and Diaconis derived expressions for optimal interval widths in fixed-interval univariate histograms and Terrell and Scott obtained corresponding results for variable-interval univariate histograms. The present paper considers the more general problem of optimal fixed and variable cell dimensions in several varieties of bivariate histograms. Optimal cell dimensions are derived, theoretically optimal histograms are constructed, and simulation studies are performed. © 1991 Academic Press, Inc.

1. INTRODUCTION

It is the purpose of the present paper to address the question of optimal parameter choice for the minimization of error in a class of histogram estimators of bivariate probability density functions. An optimization of the equal-interval historgram as a univariate probability density estimator was provided by Scott [1] and Freedman and Diaconis [2], and the possibility of improvement through interval variability was explored by Terrell and Scott [3]. Nezames [4] studied the bivariate rectangular histogram with equal bin dimensions. Scott [5] investigated nonrectangular bins and found little advantage over rectangular bins with edges parallel to the coordinate axes. There are a number of possibilities for allowing the bin dimensions to vary in order to adapt to the local properties of the densities. We have explored those many possibilities, investigating their asymptotic error rates and establishing the improvements which adaptation makes possible.

Histograms are by far the oldest and most familiar of nonparametric

Received February 28, 1990; revised April 20, 1990.

AMS 1980 subject classifications:

Key words and phrases: histogram, nonparametric density estimation, optimal class intervals.

0047-259X/91 $3.00

density estimates. Despite the availability in recent years of more sophisticated techniques, histograms continue to be far more widely used in practice than all other methods combined. This popularity in itself would justify our interest in their optimality properties. However, there are theoretical reasons for this interest as well. There is a hierarchy of density estimators which includes the histogram, frequency polygon [6], kernel [7], fourier integral estimator [8], and parametric estimators. We find faster rates of convergence as we go up the hierarchy, but at the same time, increasingly stringent smoothness requirements in order to achieve that rate. For example, the theory of histograms requires roughly that $\int (f')^2$ be reasonable, while frequency polygons require the same of $\int (f'')^2$. Further, as we go up the hierarchy, the estimates become increasingly sensitive to calibration errors in the smoothing parameters [1]. Thus, histograms are a model for resistant density estimation.

In this paper we consider only bivariate histograms. Those of higher dimension are difficult to represent graphically [9]. Furthermore, most of the characteristic difficulties already appear in two variables. We assume that the true density is known; the problem of purely data-driven calibration is not addressed here.

2. The Univariate Case

Let us first consider histogram estimators of univariate probability density functions. We begin with a random sample of size n from a population whose underlying density is $f(x)$. Let $v(x)$ be the number of sample points having values less than or equal to x. Then a natural approximation to the cumulative distribution function F is

$$\hat{F}_n = v(x)/n.$$

Similarly, a natural approximation to the probability density function f is

$$\hat{f}_k = \{v(x_k) - v(x_{k-1})\}/\{n(x_k - x_{k-1})\},$$

where the x_k are points defined by a mesh on the real line and where $v_k = v(x_k) - v(x_{k-1})$ is the number of sample points falling in the kth interval $(x_{k-1}, x_k]$. A histogram estimate is a step function $\hat{f}(x)$ of height $\hat{f}_k$ along each such interval.

The problem of optimal interval length for univariate histograms whose interval lengths remain constant throughout the domain of support was addressed by [1]. The integrated mean squared error was used there as a global measure of error.

The mean squared error (MSE) of $\hat{f}(x)$ is defined as

$$\begin{aligned}\mathrm{MSE}\{\hat{f}(x)\} &= E[\{\hat{f}(x)-f(x)\}^2] \\ &= E([\hat{f}(x)-E\{\hat{f}(x)\}]^2)+[E\{\hat{f}(x)\}-f(x)]^2 \\ &= \mathrm{var}\{\hat{f}(x)\}+\mathrm{Bias}^2\{\hat{f}(x)\}\end{aligned}$$

so that the integrated mean squared error (IMSE) becomes

$$\mathrm{IMSE}\{\hat{f}(x)\} = \int_{-\infty}^{\infty} \mathrm{var}\{\hat{f}(x)\}\,dx + \int_{-\infty}^{\infty} \mathrm{Bias}^2\{\hat{f}(x)\}\,dx.$$

Let $f'(x)$ be square Riemann integrable, with $f'(x_k)\neq 0$, and defined over the entire real line. Let h be the length of each interval and let n be the sample size so that $h(n)\to 0$ as $n\to\infty$ and $nh(n)\to\infty$ as $n\to\infty$; then the asymptotically optimal fixed interval length can be shown to be [1]

$$h^* = \left[6/n\int_{-\infty}^{\infty}\{f'(x)\}^2\,dx\right]^{1/3}.$$

When the optimal constant interval length h^* is used, the following minimal integrated mean squared error is obtained:

$$\mathrm{IMSE}^* = \frac{3}{2}6^{-1/3}n^{-2/3}\left[\int_{-\infty}^{\infty}\{f'(x)\}^2\right]^{1/3}.$$

The problem of optimality in different regions of the domain of support remains, however.

Terrell and Scott [3] and Kogure [10] addressed this issue for probability densities of one random variable. If the interval lengths are allowed to vary, that is, if h becomes a function of x, then the asymptotically optimal interval length at one point becomes

$$h^*(x_n) = [6f(x_n)/n\{f'(x_n)\}^2]^{1/3},$$

and the corresponding integrated mean squared error becomes

$$\mathrm{IMSE}^* = \frac{3}{2}6^{-1/3}n^{-2/3}\int_{-\infty}^{\infty}\{f(x)\,f'(x)\}^{2/3}\,dx.$$

Comparison of the minimum obtainable integrated mean squared error for the fixed and variable interval cases demonstrates that the minimal integrated mean squared error for the fixed interval case is always greater than or equal to that for the variable interval case. If $f(x)$ is the normal

distribution with mean equal to zero and variance equal to one, $h \simeq 3.491 n^{-1/3}$ for the fixed interval case and $h \simeq 2.469\,|x|^{-2/3} e^{x^2/6} n^{-1/3}$ for the variable interval case.

3. Histogram Estimators of Bivariate Densities

In our discussion of the one-dimensional case of histogram estimation it was apparent that there were only two possible types of mesh from which a histogram might be constructed: h could remain constant throughout the domain of support or could vary according to some criterion of optimality. However, when the concept of histogram estimation is extended to two dimensions, the number of possible grid types becomes infinite: the plane may be partitioned into sets of any shape as long as the sets are mutually exclusive and the subdivision is exhaustive. The simplest such would be a rectangular mesh with edges parallel to the coordinate axes. Scott [5] examined a number of other mesh types, some having triangular and other hexagonal shapes, but found that hexagons resulted in only slightly improved estimates at a cost of some difficulty in implementation and that regular triangles, which were also difficult to implement, resulted in worse estimates than did the rectangles. In view of the above, the mesh types in the present paper have been confined to rectangular grids having cell sides of variable length and width parallel to the coordinate axes. Note that the coordinate axes may be rotated before the procedures suggested below are applied. The best orientation would minimize our expression for optimal IMSE: the problem of finding that orientation will not be addressed here.

Let $X=(x_1, x_2)$ be a bivariate random variable with joint probability density function $f(x_1, x_2)$. If the domain of support is subdivided into rectangles of the form $(x_{1i}, x_{1i}+h_1] \times (x_{2j}, x_{2j}+h_2]$, where $h_1>0$ and $h_2>0$ are the lengths of the sides, a histogram estimator of $f(x_1, x_2)$ at the point (x_1, x_2) may be defined in analogy with the univariate case as:

$$\hat{f}(x_1, x_2) = v_{ij}(x_1, x_2)/nh_1h_2 \quad \text{for} \quad x_1 \in (x_{1i}, x_{1i}+h_1],\ x_2 \in (x_{2j}, x_{2j}+h_2]$$

where $v_{ij}(x_1, x_2)$ is the number of sample points falling in the rectangle and where h_1 and h_2 may be constant, functions of x_{1i} or x_{2j} alone, or functions of both x_{1i} and x_{2j}. The integrated mean squared error is defined as before and, in the bivariate case, becomes asymptotically:

$$\text{IMSE} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} [\{f(x_1, x_2)/nh_1h_2\} + (12)^{-1} h_1^2(\partial f/\partial x_1)^2 + (12)^{-1} h_2^2(\partial f/\partial x_2)^2]\, dx_1\, dx_2. \tag{1}$$

See Appendix and [3]. If we treat h_1 and h_2 as if they were continuously varying functions in x_1 and x_2 and take the derivative with respect to h_1 in the arbitrary bivariate direction η_{x_1} and the derivative with respect to h_2 in the arbitrary bivariate direction η_{x_2} and set the expressions equal to zero, we obtain the following conditions for $i, j = 1, 2$; $i \neq j$:

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \eta_{x_{1i}} \{(-f/nh_i^2 h_j) + 6^{-1} h_i (\partial f/\partial x_i)^2\} \, dx_1 \, dx_2 = 0. \tag{2}$$

If we solve the above equations simultaneously for h_1 and h_2 and choose the rectangular dimensions accordingly, we will obtain the minimum integrated mean squared error. Different solutions to the above equations are obtained depending upon whether h_1 and h_2 are constant, functions of only one variable or functions of both variables, each case reflecting a different type of mesh.

3.1. *Minimally Restricted "Free" Mesh*

If h_1 and h_2 are functions of both x_1 and x_2, we obtain a subdivision of the domain of support by rectangles of arbitrary dimensions. See Fig. 1a. Unfortunately, these will not necessarily be either mutually exclusive or exhaustive if we try to specify the optimal width and height in each region of the plane. Such a scheme is clearly not implementable but is rather of theoretical interest, because it provides a lower bound for the integrated mean squared error for all possible rectangular meshes whose cell sides are parallel to the coordinate axes. Since η_{x_1} and η_{x_2} are arbitrary functions of x_1 and x_2, Eq. (2) hold if and only if, for $i, j = 1, 2$; $i \neq j$:

$$f(x_1, x_2)/nh_i^2(x_1, x_2)\, h_j(x_1, x_2) = 6^{-1} h_i(x_1, x_2)(\partial f/\partial x_i)^2.$$

The solution to these equations is easily obtained for $i, j = 1, 2$; $i \neq j$,

$$h_i(x_1, x_2) = \{6f(x_1, x_2)\, |\partial f/\partial x_j|\}^{1/4} / \{n\, |\partial f/\partial x_j|^3\}^{1/4}$$

which yield an optimal integrated mean squared error

$$\text{IMSE}^* = 2 \cdot 6^{-1/2} n^{-1/2} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \{f(x_1, x_2)\}^{1/2} \times |\partial f/\partial x_1|\, |\partial f/\partial x_2|\}^{1/2} \, dx_1 \, dx_2.$$

3.2. *Fixed-Dimension "Regular" Mesh*

If h_1 and h_2 remain constant throughout the domain of support, we obtain a subdivision which is mutually exclusive, exhaustive, and easily implementable. See Fig. 1b. The histogram estimator produced by this

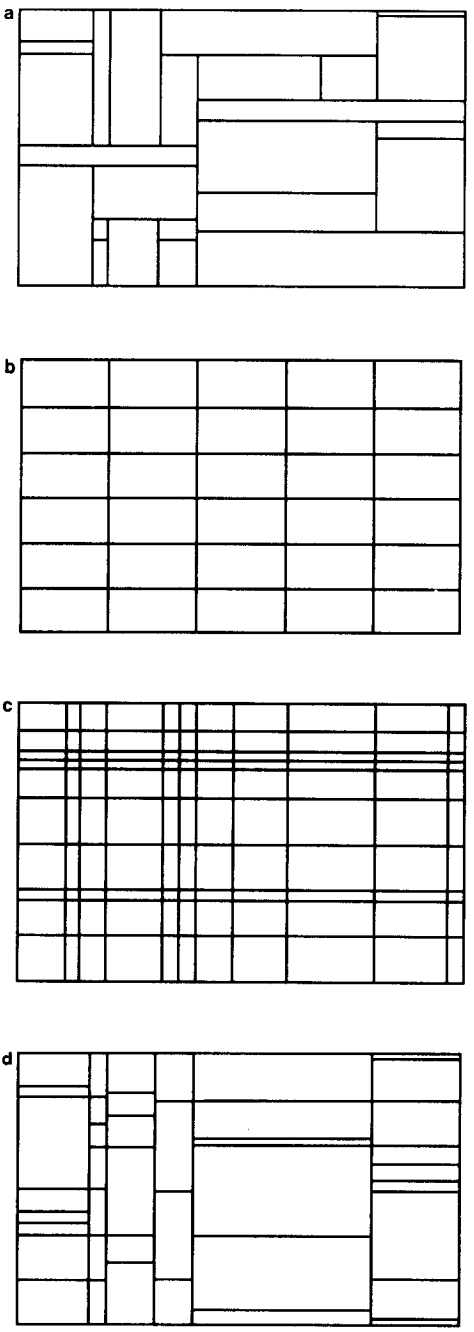

FIG. 1. (a) Free mesh; (b) regular mesh; (c) grid mesh; (d) semigrid mesh

scheme, however, is less efficient relative to that produced by the minimally restricted mesh than others which will be proposed later. When h_1 and h_2 are both constant, Eq. (2) become, for $i, j = 1, 2;\ i \neq j$,

$$1/nh_i^2 h_j = 6^{-1} h_j \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} |\partial f/\partial x_i|^2 \, dx_1 \, dx_2$$

with solution for $i, j = 1, 2;\ i \neq j$:

$$h_i = \frac{\left\{36 \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (\partial f/\partial x_j)^2 \, dx_1 \, dx_2\right\}^{1/8}}{\left[n^2 \left\{\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (\partial f/\partial x_i)^2 \, dx_1 \, dx_2\right\}^3\right]^{1/8}}.$$

Substitution of these values into (1) yields the minimal integrated mean squared error for this mesh type:

$$\text{IMSE}^* = 2 \cdot 6^{-1/2} n^{-1/2} \left\{\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (\partial f/\partial x_1)^2 \, dx_1 \, dx_2\right\}^{1/4} \times \left\{\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (\partial f/\partial x_2)^2 \, dx_1 \, dx_2\right\}^{1/2}.$$

As will be demonstrated in later sections, more efficient yet easily implementable mesh types may be designed.

3.3. *Semi-Fixed-Dimension "Semiregular" Mesh*

If h_1 is a function of x_1, and h_2 remains constant throughout the domain of support of f, we obtain a partition in which, for example, the cell widths remain constant while the lengths vary in order to accommodate changes in the form of the probability density function in different regions of its domain of support. This scheme produces a histogram estimator which is more efficient than the fixed-dimension mesh, with

$$\text{IMSE}^* = 2 \cdot 6^{-1/2} n^{-1/2} \left\{\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (\partial f/\partial x_2)^2 \, dx_1 \, dx_2\right\}^{1/4} \times \left[\int_{-\infty}^{\infty} \left\{\int_{-\infty}^{\infty} f(x_1, x_2) \, dx_2\right\}^{2/3} \left\{\int_{-\infty}^{\infty} (\partial f/\partial x_1)^2 \, dx_2\right\}^{1/3} dx_1\right]^{3/4}.$$

A similar semi-fixed-dimension mesh may be obtained if h_1 remains constant and h_2 becomes a function of x_2. The expression for h_1, $h_2(x_2)$, and the integrated mean squared error are identical to those above except that $\partial f/\partial x_1$ and $\partial f/\partial x_2$, dx_1 and dx_2 are exchanged throughout.

3.4. *Variable-Dimension "Grid" Mesh I:* $h_1(x_1)$, $h_2(x_2)$

If h_1 is a function of x_1 and h_2 is a function of x_2, we obtain a mesh which is adaptable in both dimensions to the form of the density function and which is almost as easily implemented as the fixed-dimension mesh. In several respects this type of mesh is optimal, since it combines both adaptability and ease of implementation. See Fig. 1c. Marginal histograms as well as histograms along any strip in either direction may easily be obtained. Although each such „conditional" histogram is not itself optimal, the set of all such histograms so constructed is optimal on the average. Difficulties arise, however, when an attempt is made to solve the equations deriving from (2) under the above restrictions on h_1 and h_2:

$$\{1/nh_i^2(x_i)\} \int_{-\infty}^{\infty} f(x_1, x_2)/h_j(x_j)\, dx_j = 6^{-1} h_i(x_i) \int_{-\infty}^{\infty} (\partial f/\partial x_i)^2\, dx_j.$$

If $f(x_1, x_2)$ can be written as the product of two functions, each of which is a function of only one variable, i.e., if $f(x_1, x_2) = r(x_1)\, s(x_2)$, then analytic solutions of the form

$$\begin{aligned} h_i(x_i) = 6^{1/4} n^{-1/4} \cdot \Bigg[& \int_{-\infty}^{\infty} \left\{ \int_{-\infty}^{\infty} f(x_1, x_2)\, dx_i \right\}^{2/3} \\ & \times \left\{ \int_{-\infty}^{\infty} (\partial f/\partial x_j)^2\, dx_i \right\}^{1/3} dx_j \Bigg]^{3/8} \left\{ \int_{-\infty}^{\infty} f(x_1, x_2)\, dx_j \right\}^{1/3} \\ & \times \Bigg[\int_{-\infty}^{\infty} \left\{ \int_{-\infty}^{\infty} f(x_1, x_2)\, dx_j \right\}^{2/3} \\ & \times \left\{ \int_{-\infty}^{\infty} (\partial f/\partial x_i)^2\, dx_j \right\}^{1/3} dx_i \Bigg]^{-1/8} \left\{ \int_{-\infty}^{\infty} (\partial f/\partial x_i)^2\, dx_j \right\}^{-1/3} \end{aligned}$$

may be obtained, yielding an optimal integrated mean squared error:

$$\begin{aligned} \mathrm{IMSE}^* = 2 \cdot 6^{-1/2} n^{-1/2} \Bigg[& \int_{-\infty}^{\infty} \left\{ \int_{-\infty}^{\infty} f(x_1, x_2)\, dx_1 \right\}^{2/3} \\ & \times \left\{ \int_{-\infty}^{\infty} (\partial f/\partial x_2)^2\, dx_1 \right\}^{1/3} dx_2 \Bigg]^{3/4} \\ & \times \Bigg[\int_{-\infty}^{\infty} \left\{ \int_{-\infty}^{\infty} f(x_1, x_2)\, dx_2 \right\}^{2/3} \\ & \times \left\{ \int_{-\infty}^{\infty} (\partial f/\partial x_1)^2\, dx_2 \right\}^{1/3} dx_1 \Bigg]^{3/4}. \end{aligned}$$

In the general case when $f(x_1, x_2)$ is not separable, a numerical rather than an analytic solution was found and will be treated in a later paper.

3.5. *Variable-Dimension "Semigrid" Mesh II*: $h_1(x_1)$, $h_2(x_1, x_2)$

If h_1 is a function of x_1 and h_2 is a function of both x_1 and x_2, we obtain a mesh which performs better in terms of the integrated mean squared error than the variable-dimensioned mesh described in Section 3.4 but at a cost of considerable difficulties in implementation. See Fig. 1d. An analytic solution to the equations deriving from (2) under the present restrictions is, however, readily obtainable as

$$h_1(x_1) = 6^{1/4} n^{-1/4} \left[\int_{-\infty}^{\infty} \{f(x_1, x_2)(\partial f/\partial x_2)\}^{2/3}\, dx_2 \Big/ \int_{-\infty}^{\infty} (\partial f/\partial x_1)^2\, dx_2 \right]^{3/8}$$

and

$$h_2(x_1, x_2) = 6^{1/4} n^{-1/4} \{f(x_1, x_2)\}^{1/3} (\partial f/\partial x_2)^{-2/3} \times \left[\int_{-\infty}^{\infty} (\partial f/\partial x_1)^2\, dx_2 \Big/ \int_{-\infty}^{\infty} \{f(x_1, x_2)(\partial f/\partial x_2)\}^{2/3}\, dx_2 \right]^{1/8}$$

with

$$\text{IMSE}^* = 2 \cdot 6^{-1/2} n^{-1/2} \int_{-\infty}^{\infty} \left\{ \int_{-\infty}^{\infty} (\partial f/\partial x_1)^2\, dx_2 \right\}^{1/4} \times \left[\int_{-\infty}^{\infty} \{f(x_1, x_2)(\partial f/\partial x_2)\}^{2/3}\, dx_2 \right]^{3/4} dx_1.$$

Since for each of the mesh types above we are minimizing the integrated mean squared error under increasingly tight constraints, it may be shown that the minimum integrated mean squared error becomes larges as more constraints are placed upon the mesh, in particular:

$$\text{IMSE(free)} \leqslant \text{IMSE(semigrid)} \leqslant \text{IMSE(grid)} \leqslant \text{IMSE(semiregular)} \leqslant \text{IMSE(regular)}.$$

4. Numerical Examples

Theoretically optimal histograms were constructed on the basis of the above formulations for a variety of probability density functions. Simulations were performed in which an optimal mesh was constructed for a

given density function and sample size. The empirical integrated mean squared error

$$\text{IMSE}_a = \sum_{j=1}^{n_{x_2}} \sum_{i=1}^{n_{x_1}} \int_{x_{2j}}^{x_{2j}+h_{2j}} \int_{x_{1i}}^{x_{1i}+h_{1i}} \{\hat{f}(x_{1i}, x_{2i}) - f(x_1, x_2)\}^2,$$

where n_{x_1} is the number of interval boundaries in the x_1-direction and n_{x_2} is the number of interval boundaries in the x_2-direction, was calculated for each sample and for each mesh type.

The improvements in integrated mean squared error made possible by the best variable-dimension mesh over the fixed varied from approximately 12 to 91% depending on the form of the underlying distribution.

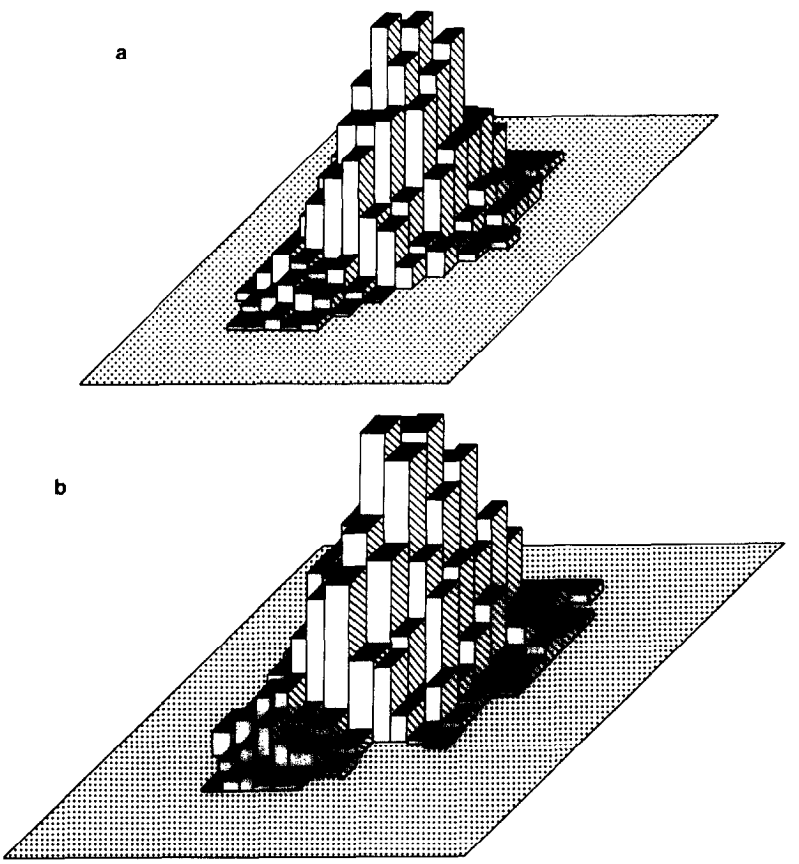

FIG. 2. (a) Regular histogram for elliptical normal with parameters $p_1 = 1$, $\rho_i = 0.8$, $\mu_{ij} = 0$, $\sigma_{ij} = 1.0$; sample size = 2000, IMSE = 7.837. (b) Semiregular histogram, IMSE = 7.811 (0.33% improvement over regular histogram). (c) Grid histogram, IMSE = 7.790 (0.60% improvement over regular histogram). (d) Semigrid histogram, IMSE = 7.023 (11.59% improvement over regular histogram).

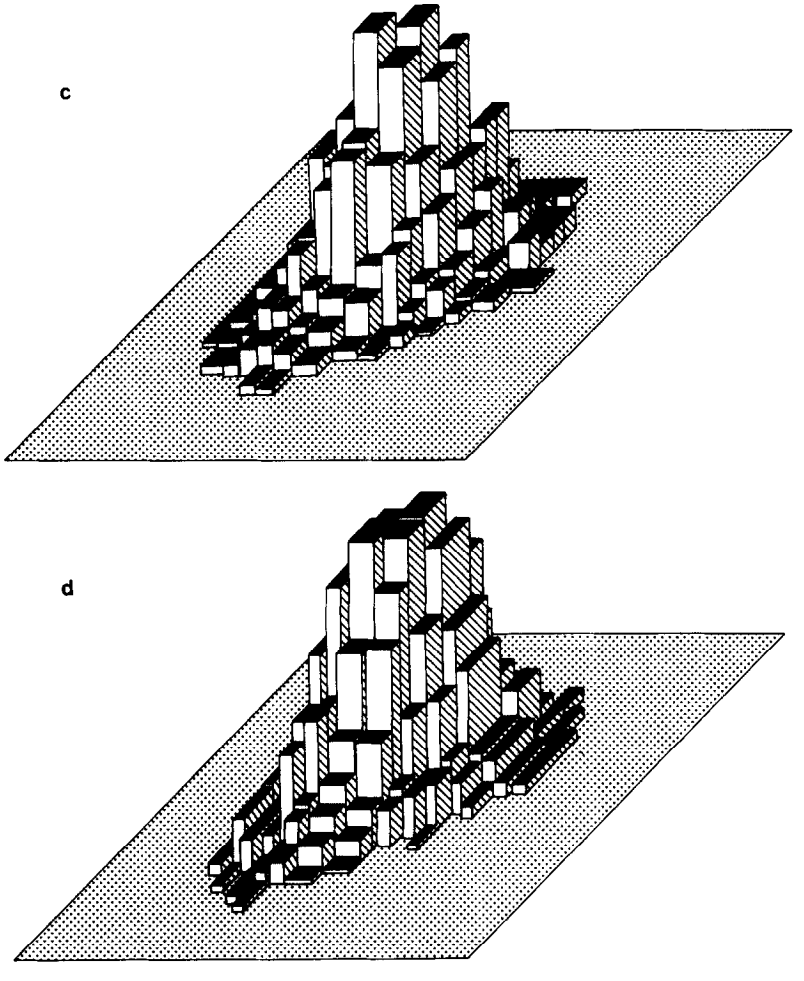

FIG. 2—*Continued*

Many cases may be constructed from a mixture of bivariate normal densities. Histograms constructed for the normal density

$$f(x_1, x_2) = \sum_{i=1}^{2} \left(0.5 p_i (\pi \sigma_{1i} \sigma_{2i} \sqrt{1-\rho_i^2})^{-1} \right.$$
$$\left. \times \exp\left[-0.5(1-\rho_i^2)^{-1} \left\{ \left(\frac{x-\mu_{1i}}{\sigma_{1i}} \right) - \left(\frac{y-\mu_{2i}}{\sigma_{2i}} \right) \right\}^2 \right] \right)$$

where $p_2 = 1 - p_1$ and having parameters $p_1 = 1$, $\rho_i = 0.8$, $\mu_{ij} = 0$, $\sigma_{ij} = 1$, for $i, j = 1, 2$, are shown in Fig. 2a through 2d. The least improvements (e.g.,

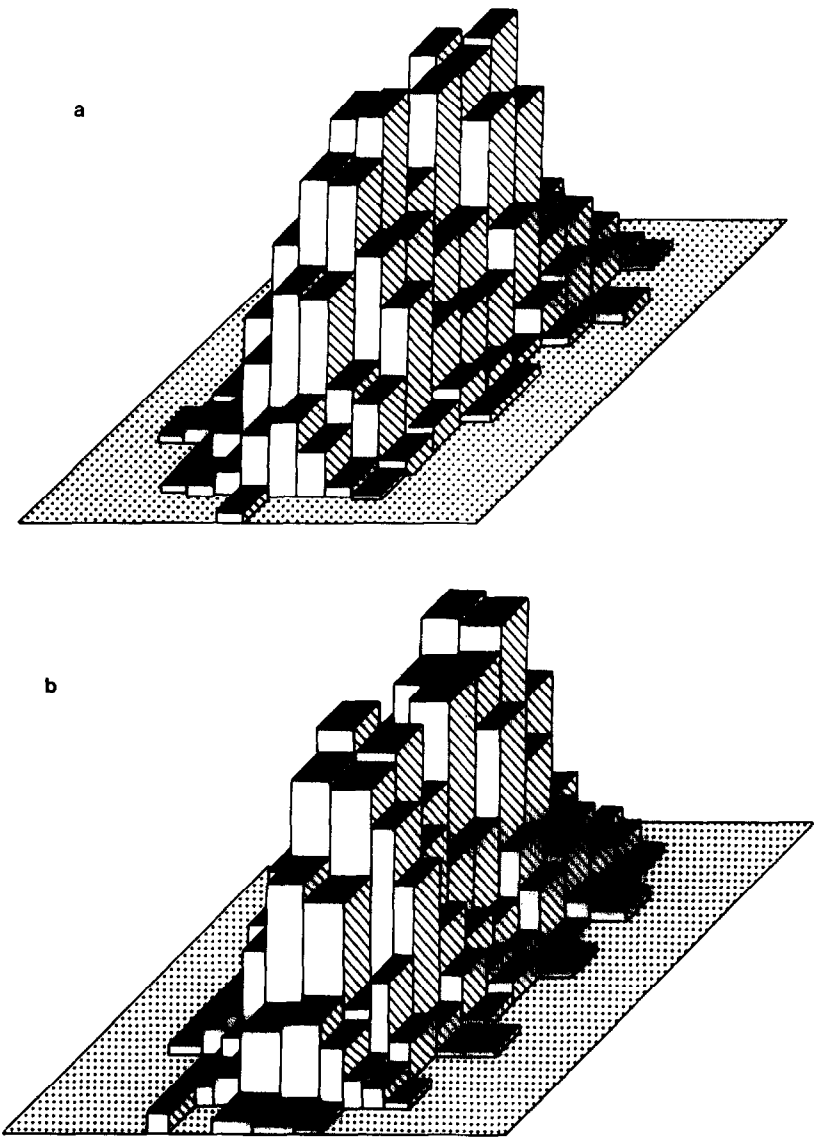

FIG. 3. (a) Regular histogram for mixed normal with parameters $p_1 = 0.5$, $\rho_i = 0$, $\mu_{11} = -1.5$, $\mu_{21} = 0$, $\mu_{12} = 1.5$, $\mu_{22} = 0$, $\sigma_{ij} = 1.0$; sample size = 2000, IMSE = 2.355. (b) Semiregular histogram, IMSE = 2.090 (12.68% improvement over regular histogram). (c) Grid histogram, IMSE = 1.915 (22.98% improvement over regular histogram). (d) Semigrid histogram, IMSE = 1.802 (30.69% improvement over regular histogram).

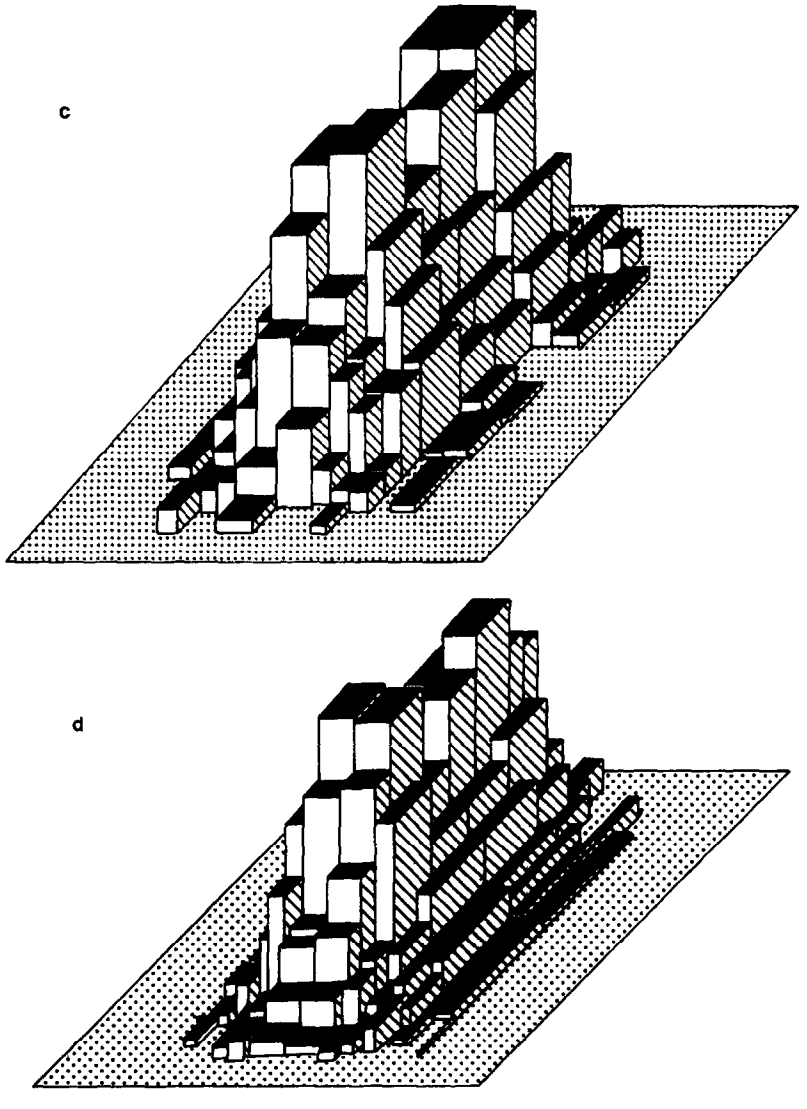

FIG. 3—*Continued*

11.59% for the semigrid histogram shown in Fig. 2) were obtained for elliptical normal densities whose major axes were nearly 45° from the coordinate axis. As mentioned earlier, adaption may be made in advance by rotation of the coordinate axes. Since adaptation must be parallel to the axes, decreasing cell size becomes the primary mode of improvement for these distributions. The effect of mesh type is substantially diminished as may be seen from the relatively small differences in integrated mean

squared errors. Rotating the axes so that the major axis of the distribution is parallel to the coordinate axes alleviates this difficulty.

Moderate improvements in approximation (10 to 40%) were obtained for the Dirichlet, unimodal normal, and mixed normal densities. Histograms constructed for the mixed normal having parameters $p_1 = 0.5$, $\rho_i = 0$, $\mu_{11} = -1.5$, $\mu_{21} = 0$, $\mu_{12} = 1.5$, $\mu_{22} = 0$, $\sigma_{ij} = 1$ are shown in Fig. 3a through 3d. A 90° rotation produces the same IMSE.

The most striking improvements in approximation (e.g., 90.95% for the semigrid over the regular in the following example) were obtained for mixed normal densities whose variances differed widely. The integrated mean squared errors for the mixed normal with parameters $p_1 = 0.5$, $\rho_i = 0$, $\mu_{11} = -1.5$, $\mu_{21} = 0$, $\mu_{12} = 1.5$, $\mu_{22} = 0$, $\sigma_{11} = \sigma_{21} = 0.2$, $\sigma_{12} = \sigma_{22} = 3.0$, where 43.05 for the regular, 30.12 for the semiregular, 25.42 for the grid, and 22.54 for the semigrid. The lower bound was 19.90 for this density.

Simulations resulted in smaller integrated mean squared errors than those predicted by the theory in Section 3. For example, the averaging of 100 simulations of the standard normal produced an empirical integrated mean squared error of 3.179 for the regular mesh, 2.992 for the semiregular, and 2.807 for the grid. The corresponding expected values were 3.643, 3.336, and 3.054. This discrepancy was found to be due to the higher order terms which do not appear in the asymptotic definition of error in Eq. (1). Similar discrepancies were reported for the univariate case [1].

Appendix

THEOREM (The integrated mean squared error for the bivariate histogram). *Let $X = (x_1, x_2)$ be a bivariate random variable with joint probability density function $f(x_1, x_2)$, the squares of whose partial derivitives are Riemann integrable; then the integrated mean squared error of $\hat{f}(x_1, x_2)$ is given by*

$$\text{IMSE} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} [\{f(x_1, x_2)/nh_1h_2\} + (12^{-1})\, h_1^2(\partial f/\partial x_1)^2 + (12^{-1})\, h_2^2(\partial f/\partial x_2)^2]\, dx_1\, dx_2.$$

Proof. If the domain of support is subdivided into rectangles of the form $(x_{1i}, x_{1i} + h_1] \times (x_{2i}, x_{2i} + h_2]$, where $h_1 > 0$ and $h_2 > 0$ are the lengths of the sides, a histogram estimator of $f(x_1, x_2)$ at the point (x_1, x_2) may be defined in analogy with the univariate case as

$$\hat{f}(x_1, x_2) = \nu(x_1, x_2)/nh_1h_2 \qquad \text{for} \quad x_1 \in (x_{1i}, x_{1i} + h_1],\ x_2 \in (x_{2i}, x_{2i} + h_2],$$

where $\nu(x_1, x_2)$ is the number of sample points falling in the rectangle and

where h_1 and h_2 may be constant, functions of x_{1i} or x_{2i} alone, or functions of both x_{1i} and x_{2i}. The integrated mean squared error is defined as before and in the bivariate case becomes:

$$\begin{aligned}\mathrm{IMSE}\{\hat{f}(x_1, x_2)\} &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} E[\{\hat{f}(x_1, x_2) - f(x_1, x_2)\}^2]\, dx_1\, dx_2 \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (E[\hat{f}(x_1, x_2) - E\{\hat{f}(x_1, x_2)\}]^2 \\ &\quad + [E\{\hat{f}(x_1, x_2)\} - f(x_1, x_2)]^2\, dx_1\, dx_2 \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} [\mathrm{Var}\{\hat{f}(x_1, x_2)\} \\ &\quad + \mathrm{Bias}^2\{\hat{f}(x_1, x_2)\}]\, dx_1\, dx_2.\end{aligned}$$

$v(x_1, x_2)$ has a binomial (n, p) distribution with p the probability that a sample point (x_1, x_2) lies int he above rectangle centered at $(x_{1i} + h_1/2, x_{2i} + h_2/2)$, so that $p = \int_{x_{2i}}^{x_{2i}+h_2} \int_{x_{1i}}^{x_{1i}+h_1} f(x_1, x_2)\, dx_1\, dx_2$, where $x_1 \in (x_{1i}, x_{1i} + h_1]$ and $x_2 \in (x_{2i}, x_{2i} + h_2]$. The variance of $\hat{f}(x_1, x_2)$ is

$$\begin{aligned}\mathrm{Var}\{\hat{f}(x_1, x_2)\} &= \mathrm{Var}\{v(x_1, x_2)/nh_1h_2\} \\ &= (1/n^2h_1^2h_2^2)\, \mathrm{Var}\{v(x_1, x_2)\} \\ &= (n/n^2h_1^2h_2^2)\, p(1-p) \\ &= (1/nh_1^2h_2^2) \int_{x_{2i}}^{x_{2i}+h_2} \int_{x_{1i}}^{x_{1i}+h_1} f(x_1, x_2)\, dx_1\, dx_2 \\ &\quad - (1/nh_1^2h_2^2) \left\{ \int_{x_{2i}}^{x_{2i}+h_2} \int_{x_{1i}}^{x_{1i}+h_1} f(x_1, x_2)\, dx_1\, dx_2 \right\} \\ &= (1/nh_1^2h_2^2) \int_{x_{2i}}^{x_{2i}+h_2} \int_{x_{1i}}^{x_{1i}+h_1} f(x_1, x_2)\, dx_1\, dx_2 \\ &\quad - (h_1h_2/nh_1^2h_2^2)\, f^2(\xi_{ij}),\end{aligned}$$

where $\xi_{ij} = (\xi_i, \xi_j)$, and $\xi_i \in (x_{1i}, x_{1i} + h_1]$, $\xi_j \in B(x_{2i}, x_{2i} + h_2]$, so that

$$\begin{aligned}&\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \mathrm{Var}\{\hat{f}(x_1, x_2)\}\, dx_1\, dx_2 \\ &\quad = (1/nh_1^2h_2^2) \sum_{j=-\infty}^{\infty} \sum_{i=-\infty}^{\infty} \int_{x_{2i}}^{x_{2i}+h_2} \int_{x_{1i}}^{x_{1i}+h_1} f(x_1, x_2)\, dx_1\, dx_2 \cdot h_1h_2 \\ &\qquad - (1/nh_1h_2) \sum_{j=-\infty}^{\infty} \sum_{i=-\infty}^{\infty} f^2(\xi_{ij}) \cdot h_1h_2\end{aligned}$$

$$= (1/nh_1h_2)\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x_1, x_2)\,dx_1\,dx_2$$

$$-(1/n)\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f^2(x_1, x_2)\,dx_1\,dx_2 + o(1/n)$$

as $n \to \infty$.

The bias may be written as

$$\begin{aligned}\text{Bias}\{\hat{f}(x_1, x_2)\} &= E\{\hat{f}(x_1, x_2) - f(x_1, x_2)\}\\ &= E\{v(x_1, x_2)/nh_1h_2\} - f(x_1, x_2)\\ &= (n/nh_1h_2)\int_{x_{2i}}^{x_{2i}+h_2}\int_{x_{1i}}^{x_{1i}+h_1} f(x_1, x_2)\,dx_1\,dx_2 - f(x_1, x_2)\\ &= (1/h_1h_2)\left[\int_{x_{2i}}^{x_{2i}+h_2}\int_{x_{1i}}^{x_{1i}+h_1}\{f(x_1, x_2) - f(x_{1i}, x_{2i})\}\,dx_1\,dx_2\right.\\ &\quad\left. - \{f(x_1, x_2) - f(x_{1i}, x_{2i})\}\right]\end{aligned}$$

and

$$\begin{aligned}\text{Bias}^2\{\hat{f}(x_1, x_2)\} &= (1/h_1^2h_2^2)\\ &\quad\times\left[\int_{x_{2i}}^{x_{2i}+h_2}\int_{x_{1i}}^{x_{1i}+h_1}\{f(x_1, x_2) - f(x_{1i}, x_{2i})\}\,dx_1\,dx_2\right]^2\\ &\quad - (2/h_1h_2)\{f(s, t) - f(x_{1i}, x_{2i})\}\\ &\quad\times\left[\int_{x_{2i}}^{x_{2i}+h_2}\int_{x_{1i}}^{x_{1i}+h_1}\{f(x_1, x_2) - fx_{1i}, x_{2i})\}\,dx_1\,dx_2\right]\\ &\quad + \{f(s, t) - f(x_{1i}, x_{2i})\}^2,\end{aligned}$$

so that the bias over one rectangle is

$$\begin{aligned}&\int_{x_{2i}}^{x_{2i}+h_2}\int_{x_{1i}}^{x_{1i}+h_1}\text{Bias}^2\{\hat{f}(x_1, x_2)\}\\ &\quad= (1/h_1h_2)\left[\int_{x_{2i}}^{x_{2i}+h_2}\int_{x_{1i}}^{x_{1i}+h_1}\{f(x_1, x_2) - f(x_{1i}, x_{2i})\}\,dx_1\,dx_2\right]^2\\ &\quad\quad - (2/h_1h_2)\left[\int_{x_{2i}}^{x_{2i}+h_2}\int_{x_{1i}}^{x_{1i}+h_1}\{f(x_1, x_2) - f(x_{1i}, x_{2i})\}\,dx_1\,dx_2\right]\\ &\quad\quad\times\left[\int_{x_{2i}}^{x_{2i}+h_2}\int_{x_{1i}}^{x_{1i}+h_1}\{f(s, t) - f(x_{1i}, x_{2i})\}\,ds\,dt\right]\\ &\quad\quad + \int_{x_{2i}}^{x_{2i}+h_2}\int_{x_{1i}}^{x_{1i}+h_1}\{f(s, t) - f(x_{1i}, x_{2i})\}^2\,ds\,dt\end{aligned}$$

$$
\begin{aligned}
&= \int_{x_{2i}}^{x_{2i}+h_2} \int_{x_{1i}}^{x_{1i}+h_1} \{f(x_1, x_2) - f(x_{1i}, x_{2i})\}^2 \, dx_1 \, dx_2 \\
&\quad - (1/h_1 h_2) \left[\int_{x_{2i}}^{x_{2i}+h_2} \int_{x_{1i}}^{x_{1i}+h_1} \{f(x_1, x_2) - f(x_{1i}, x_{2i})\} \, dx_1 \, dx_2 \right]^2 \\
&= \int_{x_{2i}}^{x_{2i}+h_2} \int_{x_{1i}}^{x_{1i}+h_1} [(x_1 - x_{1i})\{\partial f(a)/\partial x_1\} \\
&\quad + (x_2 - x_{2i})\{\partial f(b)/\partial x_2\}]^2 \, dx_1 \, dx_2 \\
&\quad - (1/h_1 h_2) \left(\int_{x_{2i}}^{x_{2i}+h_2} \int_{x_{1i}}^{x_{1i}+h_1} [(x_1 - x_{1i})\{\partial f(a)/\partial x_1\} \right. \\
&\quad \left. + (x_2 - x_{2i})\{\partial f(b)/\partial x_2\}] \, dx_1 \, dx_2 \right)^2,
\end{aligned}
$$

where $a = (x_{1i} + \xi(x_1 - x_{1i}),\ x_{2i} + \xi(x_2 - x_{2i}))$ and $b = (x_{1i} + \xi(x_1 - x_{1i}),$ $x_2 i + \xi(x_2 - x_{2i}))$ and where $0 < \xi < 1$, by the mean value theorem. Let $\xi_{x_1} = x_{1i} + \xi(x_1 - x_{1i})$ and $\xi_{x_2} = x_{2i} + \xi(x_2 - x_{2i})$. Then the above becomes

$$
\begin{aligned}
&\int_{x_{2i}}^{x_{2i}+h_2} \int_{x_{1i}}^{x_{1i}+h_1} (x_1 - x_{1i})^2 \{\partial f(\xi_{ij})/\partial x_1\}^2 \\
&\quad + 2(x_1 - x_{1i})(x_2 - x_{2i})\{\partial f(\xi_{ij})/\partial x_1\}\{\partial f(\xi_{ij})/\partial x_2\} \\
&\quad + (x_2 - x_{2i})^2 \{\partial f(\xi_{ij})/\partial x_2\}^2 \, dx_1 \, dx_2 \\
&\quad - (1/h_1 h_2) \left(\int_{x_{2i}}^{x_{2i}+h_2} \int_{x_{1i}}^{x_{1i}+h_1} [(x_1 - x_{1i})\{\partial f(\xi_{ij})/\partial x_1\} \right. \\
&\quad \left. + (x_2 - x_{2i})\{\partial f(\xi_{ij})/\partial x_2\}] \, dx_1 \, dx_2 \right)^2 \\
&= \{\partial f(\xi_{ij1})/\partial x_1\}^2 \int_{x_{2i}}^{x_{2i}+h_2} \int_{x_{1i}}^{x_{1i}+h_1} (x_1 - x_{1i})^2 \, dx_1 \, dx_2 \\
&\quad + 2\{\partial f(\xi_{ij2})/\partial x_1\}\{\partial f(\xi_{ij3})/\partial x_2\} \\
&\quad \times \int_{x_{2i}}^{x_{2i}+h_2} \int_{x_{1i}}^{x_{1i}+h_1} (x_1 - x_{1i})(x_2 - x_{2i}) \, dx_1 \, dx_2 \\
&\quad + \{\partial f(\xi_{ij4})/\partial x_2\}^2 \int_{x_{2i}}^{x_{2i}+h_2} \int_{x_{1i}}^{x_{1i}+h_1} (x_2 - x_{2i})^2 \, dx_1 \, dx_2 \\
&\quad - (1/h_1 h_2) \left[\{\partial f(\xi_{ij5})/\partial x_1\} \int_{x_{2i}}^{x_{2i}+h_2} \int_{x_{1i}}^{x_{1i}+h_1} (x_1 - x_{1i}) \, dx_1 \, dx_2 \right. \\
&\quad \left. + \{\partial f(\xi_{ij6})/\partial x_2\} \int_{x_{2i}}^{x_{2i}+h_2} \int_{x_{1i}}^{x_{1i}+h_1} (x_2 - x_{2i}) \, dx_1 \, dx_2 \right]^2
\end{aligned}
$$

by the integral form of the mean value theorem, where each ξ_{ijk}, $k = 1, ..., 6$, is a particular value of ξ_{x_1, x_2} for some x_1 and some x_2 and where

$x_1 - x_{1i} > 0$, $x_2 - x_{2i} > 0$, since $x_1 \in (x_{1i}, x_{1i} + h_1]$, $x_2 \in (x_{2i}, x_{2i} + h_2]$. The above then becomes

$$
\begin{aligned}
&\{\partial f(\xi_{ij1})/\partial x_1\}^2 h_1^3 h_2/3 + 2\{\partial f(\xi_{ij2})/\partial x_1\}\{\partial f(\xi_{ij3})/\partial x_2\} h_1^2 h_2^2/4 \\
&\quad + \{\partial f(\xi_{ij4})/\partial x_2\}^2 h_1 h_2^3/3 - (1/h_1 h_2)[\{\partial f(\xi_{ij5})/\partial x_1\} h_1^2 h_2/2 \\
&\quad + \{\partial f(\xi_{ij6})/\partial x_2\} h_1 h_2^2/2]^2 \\
&= (h_1^3 h_2/3)\{\partial f(\xi_{ij1})/\partial x_1\}^2 + (h_1^2 h_2^2/2)\{\partial f(\xi_{ij2})/\partial x_1\}\{\partial f(\xi_{ij3})/\partial x_2\} \\
&\quad + (h_1 h_2^3/3)\{\partial f(\xi_{ij4})/\partial x_2\}^2 - (1/h_1 h_2)(h_1^2 h_2^2/4)[h_1\{\partial f(\xi_{ij5})/\partial x_1\} \\
&\quad + \{h_2\, \partial f(\xi_{ij6})/\partial x_2\}]^2 \\
&= (h_1^3 h_2/3)\{\partial f(\xi_{ij1})/\partial x_1\}^2 \\
&\quad + (h_1^2 h_2^2/2)\{\partial f(\xi_{ij2})/\partial x_1\}\{\partial f(\xi_{ij3})/\partial x_2\} \\
&\quad + (h_1 h_2^3/3)\{\partial f(\xi_{ij4})/\partial x_2\}^2 - (h_1 h_2/4)[h_1^2\{\partial f(\xi_{ij5})/\partial x_1\}^2] \\
&\quad - (h_1 h_2/4)(2h_1 h_2)\{\partial f(\xi_{ij5})/\partial x_1\}\{\partial f(\xi_{ij6})/\partial x_2\} \\
&\quad - (h_1 h_2/4)\, h_2^2\{\partial f(\xi_{ij6})/\partial x_2\}^2 \\
&= (h_1^3 h_2/3)\{\partial f(\xi_{ij1})/\partial x_1\}^2 + (h_1^2 h_2^2/2)\{\partial f(\xi_{ij2})/\partial x_1\}\{\partial f(\xi_{ij3})/\partial x_2\} \\
&\quad + (h_1 h_2^3/3)\{\partial f(\xi_{ij4})/\partial x_2\}^2 - (h_1^3 h_2/4)\{\partial f(\xi_{ij5})/\partial x w_1\}^2 \\
&\quad - (h_1^2 h_2^2/2)\{\partial f(\xi_{ij5})/W\partial x_1\}\{\partial f(\xi_{ij6})/\partial x_2\} \\
&\quad - (h_1 h_2^3/4)\{\partial f(\xi_{ij6})/\partial x_2\}^2 \\
&= h_1 h_2[(h_1^2/3)\{\partial f(\xi_{ij1})/\partial x_1\}^2 - (h_1^2/4\{\partial f(\xi_{ij5})/\partial x_1\}^2] \\
&\quad + h_1 h_2(h_1 h_2/2)\{\partial f(\xi_{ij2})/\partial x_1\}\{\partial f(\xi_{ij3})/\partial x_2\} \\
&\quad - h_1 h_2[(h_1 h_2/2)\{\partial f(\xi_{ij5})/\partial x_1\}\{\partial f(\xi_{ij6})/\partial x_2\}] \\
&\quad + h_1 h_2[(h_2^2/3)\{\partial f(\xi_{ij4})/\partial x_2\}^2 - (h_2^2/4)\{\partial f(\xi_{ij6})/\partial x_2\}^2],
\end{aligned}
$$

so that

$$
\begin{aligned}
&\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathrm{Bias}^2\{\hat{f}(x_1, x_2)\} \\
&\quad = \sum_{j=-\infty}^{\infty}\sum_{i=-\infty}^{\infty} h_1 h_2[(h_1^2/3)\{\partial f(\xi_{ij1})/\partial x_1\}^2 - (h_1^2/4)\{\partial f(\xi_{ij5})/\partial x_1\}^2] \\
&\qquad + \sum_{j=-\infty}^{\infty}\sum_{i=-\infty}^{\infty} h_1 h_2[(h_1 h_2/2)\{\partial f(\xi_{ij2})/\partial x_1\}\{\partial f(\xi_{ij3})/\partial x_2\}] \\
&\qquad - (h_1 h_2/2)\{\partial f(\xi_{ij5})/\partial x_1\}\{\partial f(\xi_{ij6})/\partial x_2\} \\
&\qquad + \sum_{j=-\infty}^{\infty}\sum_{i=-\infty}^{\infty} h_1 h_2[(h_2^2/3)\{\partial f(\xi_{ij4})/\partial x_2\}^2 - (h_2^2/4)\{\partial f(\xi_{ij6})/\partial x_2\}^2].
\end{aligned}
$$

As $n \to \infty$ and $h_1, h_2 \to 0$, we have by the Riemann integrability of partial derivatives:

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \text{Bias}^2 = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (12)^{-1} h_1^2 (\partial f/\partial x_1)^2 + \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (12)^{-1} h_2^2 (\partial f/\partial x_2)^2.$$

Thus the integrated mean squared error becomes

$$\text{IMSE} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \{f(x_1, x_2)/nh_1h_2 + (12)^{-1} h_1^2 (\partial f/\partial x_1)^2 + (12)^{-1} h_2^2 (\partial f/\partial x_2)^2\} \, dx_1 \, dx_2 + o(h_1^2) + o(h_2^2) + O(1/n).$$

REFERENCES

[1] SCOTT, D. W. (1979). On optimal and ata-based histograms. *Biometrika* **66** 606–610.

[2] FREEDMAN, D., AND DIACONIS, P. (1981). On the histogram as a density estimator: L_2 theory. *Z. Wahrsch. Verw. Gebiete* **57** 453–476.

[3] TERRELL, G. R., AND SCOTT, D. W. (1983). Variable window density estimates. Presented to the *Statistical Joint Meetings, Toronto, Canada, August.*

[4] NEZAMES, D. (1980). *Bivariate Histograms.* Ph.D. dissertation, Rice University.

[5] SCOTT, D. W. (1988). A note on the choice of bivariate histogram bin shape. *J. Official Statist.* **4** 47–57.

[6] SCOTT, D. W. (1985a). Frequency polygons: Theory and applications. *J. Amer. Statist. Assoc.* **80** 348–354.

[7] ROSENBLATT, M. (1956). On some nonparametric estimates of a density function. *Ann. Math. Statist.* **27** 832–837.

[8] DAVIS, K. B. (1977). Mean integrated squared error properties of density estimators. *Ann. Statist.* **5** 530–535.

[9] SCOTT, D. W. (1985b). Average shifted histograms: Effectivegram

[10] KOGURE, A. (1987). Asymptotically optimal cells for a histogram. *Ann. Statist.* **15** 1023–1030.